IS.M.PIS21.1 Fall 2021
Sofware Engineering Project Guideline

Handed out: Tuesday, October 2, 2021

Preliminaries

In the past couple of decades modern businesses became awash with data. The question of how to best leverage this resource has been a hot topic in most forward-looking companies. The ways in which data are used to transform businesses are multiple. They range from decision support implementations where machine learning tools help businesses turn their data into actionable insights, to the use of sophisticated AI methods for solving such complex tasks like real-time communication with customers with the help of chat-bots.

For this project you should focus on machine learning techniques rather than diving into the more complex AI methods (AI will be taught to you in the spring semester). For your inspiration, below are listed some of the most common implementations of ML with specific examples. You are highly encouraged to browse through these examples, although for this course you are required to come up with your own project idea that is not readily available in any online sources.

Common implementations of ML

Decision support

Companies often use ML tools to help them make better-informed decisions. In the healthcare industry for example, decision support tools may incorporate machine learning models to guide clinicians on diagnoses and treatment options. Remember the "Diabetes" dataset we played around with in our AED course? Using it to predict whether one would become diabetic based on personal characteristics such as BMI, Age, Glucose levels, etc. would be one example of such implementation. Here is a link to ML projects executed on this dataset https://www.kaggle.com/uciml/pima-indians-diabetes-database/tasks?taskId=1163.

Customer recommendation engines

E-commerce platforms, especially large ones like Amazon or Alibaba, use recommendation engines to personalize and expedite the shopping experience; YouTube uses recommendation engines to recommend users videos that fit their tastes; another good example is the streaming entertainment service Netflix, which uses customer's viewing history, history of customers with similar entertainment interests, information about individual shows and other data points to deliver personalized recommendations to its customers – you can check out this project example for Netflix's movie recommendation engine https://www.kaggle.com/laowingkin/netflix-movie-recommendation.

Customer churn modeling

Companies utilize machine learning capabilities to analyze behaviors among existing customers to alert them to customers that are at risk of taking their business elsewhere, identify the reasons why those customers are leaving and then determine what steps to take to retain them. You can find here a project example predicting whether an existing customer of one of the top 4 wireless carriers in the US will be dropping out of their contract based on things like monthly charges, contract tenure, type of contract, etc. https://www.kaggle.com/bandiatindra/telecom-churn-prediction.

Dynamic pricing tactics

For dynamic pricing algorithms companies mine their historical pricing data along with pricing data of competitors and a host of other variables, to understand how certain dynamics impact the demand for their goods and services. Here is an example of a price prediction model, although not an example of dynamic pricing https://www.kaggle.com/goyalshalini93/car-price-prediction-linear-regression-rfe.

Market research and customer segmentation

Companies may use machine learning to predict what inventory will sell best in which of its stores based on the seasonal factors impacting a particular store, the demographics of that region and other data points. Similarly, they could use ML algorithms to better understand specific segments within their overall customer base; retailers, for instance, use the technology to gain insights into the buying patterns of specific groups of shoppers based on things like age or income or education level, etc. to better target their customers through advertisements and special offers. Here is an example of a customer segmentation project https://www.kaggle.com/fabiendaniel/customer-segmentation.

Fraud detection

For fraud detection companies use available data to understand an individual customer's typical behavior, such as when and where the customer uses a credit card. Machine learning can take that information, as well as other data sets, to accurately determine in mere milliseconds which transactions fall within the normal range and are therefore legitimate versus which transactions are outside expected norms and therefore are likely fraudulent. Here is an example of such a project https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets.

Getting Started

For starters, contact your university and company mentors because no one will do that for you. Their task is to provide guidance and help you with your questions. A good idea for the teams is to choose a team leader who would interface with the team's mentor and would create a favourable working environment. Define communication channels between team members and your mentor and find tools to manage your tasks. Be proactive and try to include every team member in the work process. Base any communication between yourselves, or the mentor, on respect.

The Task: Midterm 1

For the first midterm you will have fulfilled the following three stages of your data science project.

1. Crystallize your project's topic and objectives;

2. Collect the data and understand the data fields in the dataset;

3. Scrub and Explore your data;

Choosing a topic

The purpose of this project is to use ML models to either learn something about your data or to predict a particular dependent variable of interest.

There are at least three keys to choosing a topic. First, try to pick a field that you find interesting and that you know something about. If you enjoy working on your project, the hours involved will seem to fly by. In addition, if you know something about your subject, you'll be more likely to make correct specification choices and to notice subtle indications of data errors or theoretical problems. A second key is to make sure that data are available with a reasonable sample (we suggest at least 250 observations) or if not, that you can scrape this data easily off the internet. Nothing is more frustrating than searching through data source after data source in search of numbers for your dependent variable or one of your independent variables, so before you lock yourself into a topic, see if the data are there. The final key is to make sure that there is some substance to your topic. Try to avoid topics that are purely descriptive or virtually tautological in nature. Instead, look for topics that address an inherently interesting behavioral question or choice.

Perhaps the best place to look for ideas for topics is the internet. ML has been getting a lot of attention in the past years and there are plenty of online sources that discuss diverse implementations of ML and AI (browse for example Kaggle or github's website for ML projects).

If you get stuck for a topic, go directly to the data sources themselves. That is, instead of thinking of a topic and then seeing if the data are available, look over what data are available and see if they help you generate ideas for topics. Quite often, a reference will have data not only for a dependent variable but also for most of the relevant independent variables all in one place, minimizing time spent collecting data.

Collecting Data

Before any quantitative analysis can be done, the data must be collected, organized, and entered into a computer. Usually, this is a time-consuming and frustrating task because of the difficulty of finding data, the existence of definitional differences between theoretical variables and their empirical counterparts, and the high probability of data entry errors or data transmission errors. In general, though, time spent thinking about and collecting the data is well spent, since a researcher who knows the data sources and definitions is much less likely to make mistakes using or interpreting model results run on those data.

Here are some sources where you can find data with hypothetical examples of project ideas:

- National Centers for Environmental Information world's largest provider of weather and climate data.
- World Happiness Report 2021 for example you may want to explore what makes the world's happiest countries so happy
- NASA If you're interested in space and earth science.
- US Census Datasets about the people and economy of the United States.
- FBI Crime Data Explorer (CDE) for crime data collected by more than 18,000 law enforcement agencies.
- World Health Organization COVID-19 Dashboard Find the latest coronavirus numbers by country or WHO region.
- Latest Netflix Data This Kaggle dataset (updated in April 2021) includes movie data broken down into 26 attributes.
- Google Books Ngram From Google Books Ngram you can explore phrase trends in books published from 1960 to 2015.
- NYC Open Data publicly available datasets for New York City. Topics like the Central Park squirrel population, motor vehicle collisions, etc.
- Yelp Open Dataset Collection of Yelp user reviews, check ins, and business attributes.

Other websites where you can find rich datasets:

- CDC Wonder
- Data.gov
- World Bank
- Data.world
- /r/datasets

Advanced data sources

While you'll find no shortage of excellent (and free) public data sets on the internet, you might want to find and scrape your own data as well. This can be useful if for example you decided to go with the same or a similar topic you found on one of the websites discussing ML implementations. In order for it not to be plagiarism, you should work on a dataset that has not been analyzed before and web-scraping is a great way to create your own dataset. Additionally knowing how to scrape web data means you can find and use data sets that match your interests, regardless of whether or not they've already been compiled. If you know some Python, you can use tools like Beautiful Soup or Scrapy to crawl the web for interesting data. If you're unsure where to start, here are some websites with interesting data options to inspire your project:

- Reddit
- Wikipedia
- Job portals

Another possibility is to scrape public review sites and social media platforms, where people are likely to offer public opinions on various subjects. And perform sentiments analysis. To do so, you may have to dip into AI tools since Neural Networks and Recurrent Neural Networks are favored for NLP tasks.

To get started exploring what people feel about a certain topic, you can check out sites like:

- Amazon (product reviews)
- Rotten Tomato (movie reviews)
- Facebook
- Twitter
- News sites

Data cleaning and exploring

Cleaning

In this process, you need to convert the data from one format to another and consolidate everything into one standardized format across all data. For example, if your data is stored in multiple CSV files, then you will consolidate these CSV data into a single repository, so that you can process and analyze it.

To clean the data you will have to:

- Check that the data types are correctly specified;
- See if any data fields require special encoding (e.g. create dummies, group variables into ranges or bins, etc.);
- Get rid of duplicates;
- Decide how to deal with the outliers, errors and missing values.

To correctly deal with the last one, you will have to know how the data was collected. The good news is that we have a direct connection with the entity that collects the data, thus we can always find out necessary details about the collection process that will help you decide how to deal with the missing values, errors, and outliers. Remember: just because an outlier exists, doesn't mean it is incorrect. One needs to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it. Otherwise do not remove.

The data cleaning stages and the data exploration will have to be done to some extent concomitantly. For example you will want to look at histograms and box plots to learn about outliers. Before this however you will have to deal with missing values and duplicates. Cross-classified scatter plots, heat maps and least square lines are not useful until you have fully cleaned your data. On the other hand you should compute the summary statistics of the relevant data fields to learn about the missing values and errors.

Exploring

After cleaning the dataset you will have to fulfill an integral exploration of your data. You will look into summary statistics (means, medians, modes, interquartile ranges, skewness, etc.) and develop a good understanding of variables' distributions by analyzing their box plots and histograms, even if you have done so before having the data scrubbed. You need to have a comprehensive picture of the dataset that you will be fitting the models on, that is the dataset after it has been scrubbed. You should be able to analyze now the most promising correlations by studying the dataset's cross-classified scatter plots, heat maps and/or least squares lines. Please feel free to use any additional methods to explora data that you may have learned in your class on data visualization.

In the end of step 3 you should go through a validation and a QA process where you will try to answer the following questions:

- oes the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your intuition, or bring any insight to light?
- Can you find trends in the data to help you form/change your expectations?
- If not, is that because of a data quality issue?

During this stage we should have at least one, possibly two, meetings, where you will report and we will discuss your findings from the data exploration stage as well as the techniques you employed while scrubbing the data. You should consider reshaping some of the arguments and expectations created in step 2 given your recent findings from the exploration stage.

The Task: Midterm 2

For the second midterm you'll be working on:

1. Feature selection and/or dimensionality reduction;
2. Fitting at least two different models to make the predictions;
3. Selecting best performing model and boosting the model (if possible);
4. Interpreting the data.

This is the most interesting stage of the data science project lifecycle. As many people would call it “where the magic happens”. Once again, before reaching to this stage, bear in mind that the scrubbing and exploring stage is crucial in order for this process to make sense. So take your time on those stages instead of jumping right to this process.

Feature selection and/or dimensionality reduction

One of the first things you need to do in modeling data, is to reduce the dimensionality of your data set. Not all your features or values are essential to predicting your model. Hence, what you need to do is to select the relevant ones that will contribute to the prediction of results you are looking for.

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.

You are not required to implement any specific algorithms for feature selection unless you have to. When selecting the features that go into your model however you should have a good understanding of the reason why you chose those features (you should have a sound theoretical reason why you chose those variables possibly reinforced by statistical metrics such as correlations, p-values, etc.) - which is the most basic type of feature selection.

You should nevertheless test different methods of dimensionality reduction. You would have studied by now a number of methods for dimensionality reduction. This is where you should try to implement what you have learned into practice. You may in the end decide to include all the features contained in your dataset, yet you should compare the predictive power of your model after employing different methods of dimensionality reduction to demonstrate that including all features is the best solution.

Modeling

This is indeed the most interesting part of the project. In reality it will happen to a certain extent concomitantly with the first step of dimensionality reduction.

You should chose, two - three alternative models depending on the type of dependent variable you are dealing with and the problem that you are trying to solve. By this stage you would have probably finished the ML class and should be familiar with most usual modeling techniques. If you are trying to predict a number you may look into regression models. Here we have multinomial regressions, polynomial regressions, support vector regressions, random forest regressions and others. If your problem is of clustering nature you have options from the simplest k-means to hierarchical models, to more sophisticated algorithms like HDBScan. The third most common problem would be a classification problem and here you can use Logistic regressions, K-nearest neighbours (K-NN), Support Vector Machine (SVM) algorithms, Kernel SVM, Naïve Bayes, Random Forest and other.

You should schedule some meeting(s) with your adviser to report on and discuss about the types of models you decided to fit and your reasoning behind your decision.

Interpreting

This is the one before last and extremley important step of the project, that is, interpreting data. Interpreting data basically refers to the presentation of your data, delivering the results in such a way that is able to answer the business questions you asked when you first started the project, together with the actionable insights that is found through data science.

Actionable insights is a key outcome in which we can show that data science can bring about predictive analytics and later on prescriptive analytics, in which we can learn how to repeat a good outcome, or prevent a negative outcome.

In this process, technical skills only is not sufficient. One very important skill you need is to be able to tell a very clear and actionable story. If your presentation does not trigger actions in your hypothetical audience, it means that your communication was not efficient.

The Task: Exam

For the exam you'll be working on:

1. Finalizing your project's report;
2. Preparing each other for the exam.

After receiving feedback from the second Midterm, it is now time to finalize your project. Make sure to follow your mentor's instructions on how to enhance your application, what needs to be added and what needs to be re-done in the report. After applying all the needed changes, you will upload the final version of the report on Else, together with a link to a publicly available repository with your application's code. After this, your mentor will mark your final project (called in the next chapters *mentor_grade*).

For the exam itself, expect an individual interview with your mentors. Brush up on theoretical concepts learned during the project and make sure that every team member knows about what other team members did in the project. The fact that you didn't do that part of the project is not a valid excuse.

Reporting

Your deliverables for this project are the team presentations, a team report and the models you've built. At presentations, try to explain what did you do, how did you do it and why did you choose to do it like you did. Bring arguments, cover only the important stuff and keep your presentations interesting. Presentations should mimic your report structure.

The credo for the project's report is ”High quality content, with low levels of bullshit”. There are no page limits, neither upper nor lower bound, so don't fill your reports with water – better use it to plant some trees. For report structure, take inspiration from the previous sections where the tasks were described. Don't forget about about a title page and a references section. Not stating your sources is considered plagiarism and *will not be tolerated.*

This project requires the use of LaTeX for text editing (a safe choice is Overleaf). Also, consider using Grammarly – it's a free tool that will help you fix not only spelling but also bad grammar and awkward sentence formulations.

Evaluation

The final grade for this project will be calculated as follows:

$$\begin{aligned} final\_grade = {} & 20\% \times individual\_midterm1\_grade \\ & + 20\% \times individual\_midterm2\_grade \\ & + 20\% \times mentor\_grade \\ & + 40\% \times final\_exam\_grade \end{aligned}$$

Both the mentor grade and final exam grade are given to you individually. However midterm presentations are going to be team presentations for which you will receive a team grade. Individual midterm grades will then be calculated as follows:

$$individual\_midterm\_grade = team\_midterm\_grade + individual\_correction$$

For both Midterm 1 and Midterm 2 presentations, each team must submit a list of corrections in which they appreciate each member's contribution to the project. Below are 2 examples of such a list:

Team 1		**Team 2**	
Andronic G.	0	Popa A.	0
Popova Z.	0	Cusnir A.	+1
Zamfir I.	0	Petrov O.	+1
		Ivanov I.	−2

You will submit 1 list of corrections per team. The possible correction values are `[-2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5]`, and their sum should always be equal to `0`. After the presentation, a student's individual midterm grade will be calculated by the formula above. If e.g. both Team 1 and Team 2 have received an `8` for their presentations, their individual marks will be as follows:

Team 1		**Team 2**	
Andronic G.	$8 + (0) = 8$	Popa A.	$8 + (0) = 8$
Popova Z.	$8 + (0) = 8$	Cusnir A.	$8 + (+1) = 9$
Zamfir I.	$8 + (0) = 8$	Petrov O.	$8 + (+1) = 9$
		Ivanov I.	$8 + (-2) = 6$

As such, these corrections could be used to appreciate teammates that had a significant input to the project during the half-semester. Conversely, the corrections could be used to penalize teammates which did not prove helpful in the team during the project.

Good Luck!